Improving Phylogenetic Input Data with Phylogenetic Focusing

Christopher J. Gonzalez and David C. Plachetzki

Department of Molecular, Cellular, Biomedical Science,
University of New Hampshire, Durham, NH 03824

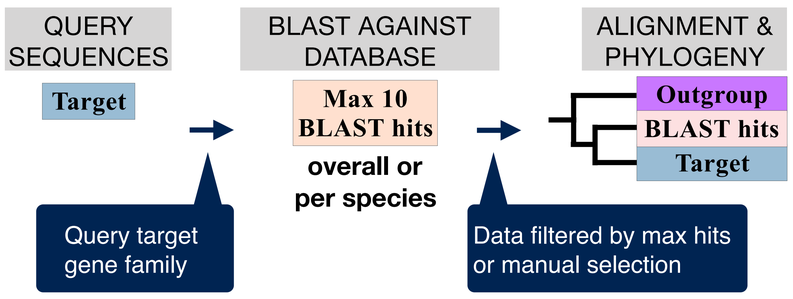
Conceptual Background

- Mass sequence data provides unmatched opportunity for studies on gene family evolution & phylogenetics.
- Not all sequence data is relevant for every analysis, and efficiency requires limits on the volume of data used.
- How to filter input sequences for phylogenetics is crucial.

Input Data Filtering Goals

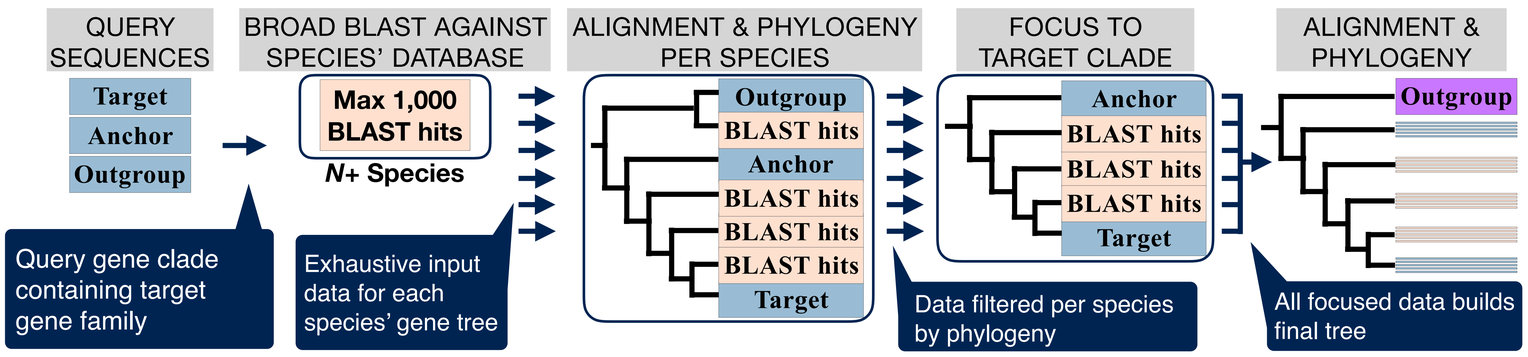
- limit input data size for computational feasibility
- Exclude irrelevant sequences
- Include relevant sequences

Phylogenetic Approaches to Input Data Filtering



General Approach (simplified)

- Many variations; most filter input data via BLAST hit limits and/or manual curation.
- One alignment and phylogeny made from all filtered data.
- Often a good starting point, but BLAST hits can include irrelevant data, while hit limit may exclude relevant data.



PhyFocus Approach

- Filters input data by assessing each species' gene phylogeny; phylogenies thus dictate what input data is relevant or irrelevant.
- Per species approach reduces data volume while eliminating need for arbitrary BLAST hit limits or manual selection bias.
- Requires extensive species sequence data and some prior phylogenetic knowledge for query gene clade.

Approaches Example: Chemosensory Genes

Query Target Seqs (*Homo* & *Danio*):

- Taste 1 Receptors (T1Rs)
- Calcium Receptor (CasR)

BLAST Database:

- 43 Deuterostome peptide datasets

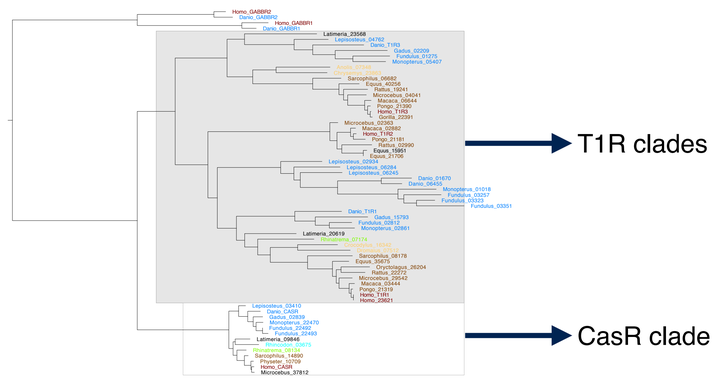


Figure 1. General Approach ML phylogeny

- Only target clades identified
- Limited species diversity

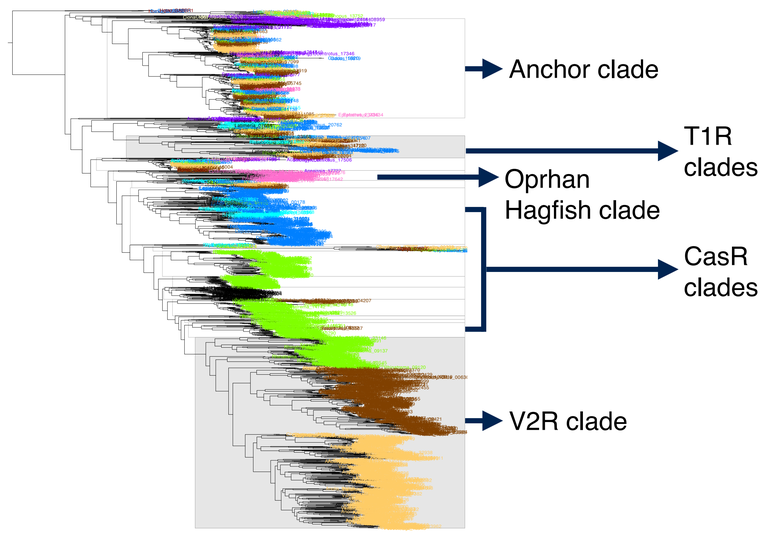


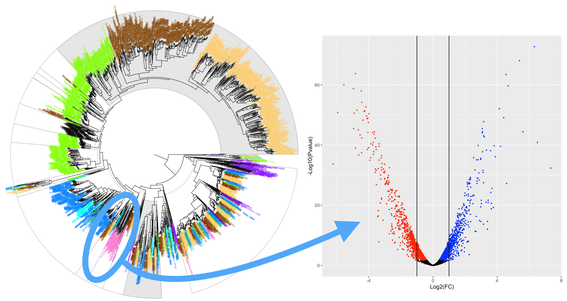
Figure 2. PhyFocus ML phylogeny

- T1R, CasR, & V2R clades identified
- Greater species diversity, resolution on orphan clades

Additional PhyFocus Applications

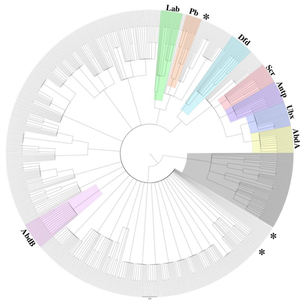
Candidate Gene Identification

- candidates for hagfish chemosensory genes (left) can be targeted in RNAseq of sensory tissues (right)



Rigorous Homolog Identification

- Support annotation of mayfly Hox gene homologs via placement in ANTP-class gene phylogeny[1]



Getting Started with PhyFocus

Required Dependencies

- AWK
- Python3
- R
- Perl
- BLAST+
- CD-HIT
- MAFFT
- IQTREE
- HMMER

Required Files

- Query FASTA
- Query alignment FASTA
- Species' peptide FASTAs
- Query header names CSV

PhyFocus Access

- At github.com/C-gonz

Example Specifications

- CPUs: 24
- RAM: 125GB
- Figure 2 runtime: 7-10 days

Acknowledgements

- We would like to thank UNH MCBS for supporting this work, and fellow MCBS lab members for their feedback.
- Additional thanks for guidance and technical support from the UNH Hubbard Center for Genome Studies.

Works Cited

1. Gonzalez, C. J., Hildebrandt, T. R., & O'Donnell, B. (2022). Characterizing Hox genes in mayflies (Ephemeroptera), with *Hexagenia limbata* as a new mayfly model. EvoDevo, 13(1), 1-17.